Content, Social, and Metacognitive Statements: an empirical study comparing human-human and human-computer tutorial dialogue

Myroslava O. Dzikovska[1], Natalie B. Steinhauser[2], Johanna D. Moore[1], Gwendolyn E. Campbell[2], Katherine M. Harrison[3], and Leanne S. Taylor[4]*

[1] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
{m.dzikovska,j.moore}@ed.ac.uk
[2] Naval Air Warfare Center Training Systems Division, Orlando, FL, USA
{natalie.steinhauser,gwendolyn.campbell}@navy.mil
[3] Kaegan Corporation, 12000 Research Parkway, Orlando, FL 32826-2944
Katherine.M.Harrison.ctr@navy.mil
[4] University of Central Florida, 4000 Central Florida Blvd. Orlando, FL 32816
Leanne.Taylor.ctr@navy.mil

Abstract. We present a study which compares human-human computer-mediated tutoring with two computer tutoring systems based on the same materials but differing in the type of feedback they provide. Our results show that there are significant differences in interaction style between human-human and human-computer tutoring, as well as between the two computer tutors, and that different dialogue characteristics predict learning gain in different conditions. We show that there are significant differences in the non-content statements that students make to human and computer tutors, but also to different types of computer tutors. These differences also affect which factors are correlated with learning gain and user satisfaction. We argue that ITS designers should pay particular attention to strategies for dealing with negative social and metacognitive statements, and also conduct further research on how interaction style affects human-computer tutoring.

1 Introduction

Intelligent Tutoring Systems (ITS) are often used as part of technology-enhanced curricula, either on their own to help students practice skills [1] or in a wider context by providing support for exercises in an e-learning environment [2]. One approach to creating ITSs is to model them after a human tutor because human tutoring combined with classroom teaching has been said to be the most effective

* This work has been supported in part by US Office of Naval Research grants N000141010085 and N0001410WX20278. We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez and five former Research Associates who worked on this project, Charles Callaway, Elaine Farrow, Leslie Butler, Lisa Durrance, and Cheryl Johnson.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Content, Social, and Metacognitive Statements: an empirical study comparing human-human and human-computer tutorial dialogue**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **School of Informatics,University of Edinburgh, Edinburgh,,United Kingdom, ,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We present a study which compares human-human computer- mediated tutoring with two computer tutoring systems based on the same materials but differing in the type of feedback they provide. Our results show that there are significant differences in interaction style be- tween human-human and human-computer tutoring, as well as between the two computer tutors, and that different dialogue characteristics pre- dict learning gain in different conditions. We show that there are sig- nificant differences in the non-content statements that students make to human and computer tutors, but also to different types of computer tutors. These differences also affect which factors are correlated with learning gain and user satisfaction. We argue that ITS designers should pay particular attention to strategies for dealing with negative social and metacognitive statements, and also conduct further research on how interaction style affects human-computer tutoring.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**16**	

form of instruction [3]. However, there remains a significant gap between the capabilities of a human tutor and the capabilities of even the best existing ITS, so the direct comparison between human-human and human-computer tutoring interactions may be difficult.

One of the key capabilities of human tutors is the ability to engage in natural language dialogue with students, providing scaffolding, explanations and motivational prompts. There are now a number of ITSs that engage in some form of natural language interaction with a student. These systems vary widely with respect to the type of natural language input they support and the kinds of feedback they provide. Student input may be restricted to short answers to questions (single words or phrases) [4, 5][5], support a small set of longer sentences [6–10], or attempt to interpret extended answers to "Why" questions [8, 11, 12]. The feedback may be completely hand-authored [4–6, 8], or generated automatically based on the system's internal representation [7, 9, 10, 12].

This large space of possibilities can create very different interaction styles. For example, pre-authored feedback can provide the greatest flexibility in the form of the instructional strategies employed, but is mostly possible in systems that rely on short-answer questions or limited domain size. This greatly constrains the language that the system can understand, and means that interaction with such systems is not really similar to interaction with human tutors, even if the feedback is authored by humans. Pre-authored feedback is also difficult to adapt to the student model or past interaction history, sometimes leading to redundancy and student confusion [13]. In contrast, automatically generated feedback is more flexible and together with unrestricted language input, it may make the interaction more similar to human-human interaction. But such systems are currently more prone to errors, both in interpreting and in generating feedback, and may have to use a more constrained set of strategies due to limitations of existing reasoning and natural language generation techniques.

Given the large space of possibilities and tradeoffs between them, it is not clear which of the techniques used by human tutors will be effective in human-computer interaction. Moreover, implementing such techniques in computer systems relies on the assumption that students will react to the computer tutors in the same way they do to human tutors, and therefore that the same tutoring strategies will promote learning. To understand this problem better, we have performed a controlled experiment that compares three conditions: an ITS that asks students to explain their reasoning in their own words, and provides feedback targeted to the errors (if any), a version of the same system that gives away correct answers without specific feedback, and a human tutor using the same instructional materials in a computer-mediated learning environment.

We focus on two aspects of the interaction: the role of student-produced content, since it has been shown to predict learning gain with both human and computer tutors [14, 15], and social and metacognitive utterances produced by students, since these are related to affective states, and have also been shown to

[5] The Why2-Atlas tutor is capable of interpreting long student essays, but restricts student input to short-answer questions in dialogue interactions.

predict learning gain [16]. Our results indicate that there are significant differences in human-human and human-computer interaction, and that in computer tutoring, different factors predict learning gain depending on the type of feedback given. This may have implications for devising appropriate feedback in computer-based learning environments.

This paper is organized as follows. In Section 2 we review the literature on differences between human-human and human-computer interaction, and factors which predict learning gain in tutorial dialogue. In Section 3 we discuss our human-human tutoring study, followed by the human-computer study in Section 4. The coding scheme used to analyze the data is described in Section 5. Section 6 discusses results and how the interaction and learning outcome differ in human-human and human-computer tutoring. We discuss the potential implications for technology-enhanced learning systems and future work in Section 7.

2 Background

Current research on how people respond to computers and computer entities, in comparison to humans, has produced mixed results. The work of Reeves & Nass [17] shows that people treat computers as social actors, i.e. they unconsciously follow rules of social relationships when interacting with media, and display emotions common in human-human relationships such as politeness or anger. Further studies demonstrated that people often respond to virtual humans similarly to how they respond to real people [18, 19].

In contrast, more recent research using ITSs shows that students change the language they use in dialogue depending on whether they are interacting with humans or computers. When talking to a computer, students who were led to believe they were conversing with a human used more words and conversed longer than did students who knew they were talking to a machine [20]. Students also provided more explanations and longer turns when they thought they were conversing with a human versus a computer, even though they were conversing with a computer in both situations [21].

The use of natural language interaction has long been hypothesized as one of the factors that can contribute to improved learning outcomes with computer systems. There is a body of research suggesting that the kind of language the students are producing, both with human and with computer tutors, is important as well: a higher percentage of contentful talk is correlated with higher learning gain [14, 15], and getting students to self-explain improves learning [22]. Yet studies comparing human tutors, computer tutors and carefully designed reading materials failed to demonstrate significant differences in learning gains [23], except when students were being taught by human tutors on content in their Zone of Proximal Development [24].

An open question when comparing human-human and human-computer tutoring results, however, is how much they are impacted by differences in human-human and human-computer interaction. For example, the systems in [23] all relied on short-answer questions, with rigidly defined dialogue structures, resulting in a very different interaction style than that observed in the human tutoring.

The goal of the experiment discussed in this paper is to provide a more controlled comparison between human-human and human-computer tutoring. For this purpose, we developed a computer-mediated learning environment where human tutors helped students work through a set of exercises. This framework fixed the choice of exercises, but let the tutors use their own strategies in providing feedback to students. Based on the data we collected from the human-human interaction, we developed a tutorial dialogue system that replaces the human tutor's feedback with automatically generated computer feedback, using natural language processing techniques to analyze student explanations. While the system does not achieve human competency in terms of interpreting natural language, the ability to accept less restricted input provides a closer comparison to the kinds of dialogue we have seen in human-human interaction, and therefore gives us a better opportunity to investigate whether the same factors are related to learning gain in the two situations.

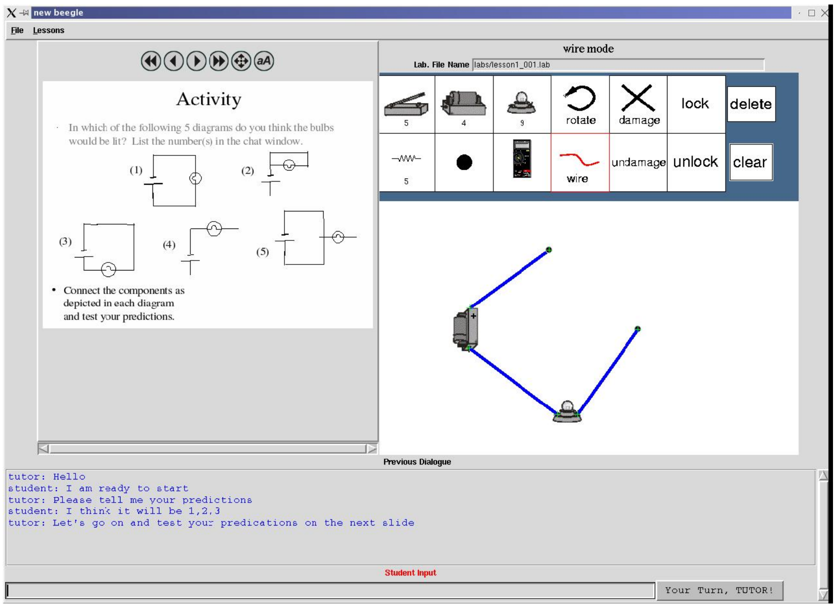
Our initial studies comparing human-human and human-computer data demonstrated that two types of variables are important for learning. First, we found evidence that students who produce a higher percentage of content words learn more [15] (see also [14] for a similar result in another tutoring system). Second, we showed that different forms of non-content talk (social and metacognitive statements) are correlated with learning in human-human dialogue and in human-computer dialogue [16, 25]. However, different factors are (positively or negatively) correlated with learning gain in human-computer versus human-human dialogue.

In this paper we extend our data analysis to explicitly compare three conditions: human-human tutoring, a computer tutor with adaptive feedback, and a computer tutor which gives away the answers without providing specific feedback. We examine the correlations between both content and non-content student statements and learning gain, and discuss implications for tutoring system design.

3 Human-Human Tutoring Study

3.1 Data collection environment

We constructed a curriculum incorporating lessons in basic electricity and electronics for use in computer-mediated instruction, including reading materials, interactive exercises with a circuit simulator, and general discussion questions. The curriculum covered topics including open and closed paths, voltage reading between components and positive and negative terminals, series and parallel configurations, and finding faults in a circuit with a multimeter. The students were asked to read slides, interact with a circuit simulator, and explain their answers, with questions like "explain why bulb A was on when switch C was open." The exercises also asked some high-level questions such as "What are the conditions for a bulb to light?".



Lab. File Name labs/lesson1_001.lab

Activity

- In which of the following 5 diagrams do you think the bulbs would be lit? List the number(s) in the chat window.
- Connect the components as depicted in each diagram and test your predictions.

```
tutor: Hello
student: I am ready to start
tutor: Please tell me your predictions
student: I think it will be 1,2,3
tutor: Let's go on and test your predications on the next slide
```

Fig. 1. Participant screen for human-human tutoring

Figure 1 shows a screenshot of the learning environment that the participants interacted with during the study. The screen is divided into three sections. The top left-hand section displays slides which deliver core lesson material including educational text, activities, and discussion questions. The participants were able to continue through the lesson slides at their own pace. The top right-hand section contains a circuit simulator which allows participants to construct and manipulate circuits as a supplement to the material in the slides. The bottom section is the chat window where the participants and tutor converse by typing.

The tutor and student were not co-located, however the tutor did have the ability to observe the student's learning environment and interact with the student through a chat window. The tutor gave feedback, technical assistance and encouragement as appropriate. Participants directed their answers, comments, and/or questions to the tutor throughout the curriculum. Each session with a system lasted approximately 4 hours.

3.2 Procedure

After completing informed consent paperwork, participants filled out a demographic questionnaire and took a pre-test consisting of 38 multiple choice questions. The participants were then introduced to their tutor and given a brief demonstration of how to operate the learning environment. The students spent the majority of the experimental session working through the lesson material and building circuits. At the conclusion of the experiment, participants completed a post-test which included 21 multiple choice questions and a satisfaction questionnaire. They were then debriefed and excused.

3.3 Corpus

Thirty undergraduate students from a large South-East US university participated in the study. The participants had little prior knowledge of the domain, and were paid for their time. They were randomly distributed across three experienced tutors. The average age of the participants was 22.4 years ($SD = 5.0$) and exactly half of them were female. The records of the sessions comprise the corpus for this study, which contains 8,100 dialogue turns taken by the student and tutor, and 56,000 words (including punctuation).

4 Human-Computer Tutoring Study

4.1 Data collection environment

To the extent possible, the same curriculum was used in the human-human study and in the computer tutoring system (Beetle II). Small changes were made to the curriculum so that the computer would be able to understand student responses more easily. For example, multi-part questions were sometimes broken into series of single questions (e.g., the question "Explain your reasoning for those 5 diagrams" would be replaced with questions "Explain your reasoning for the circuit in diagram 1", "Explain your reasoning for the circuit in diagram 2" and so on). The user interface had the same functionality as the human-human tutoring environment.

Within the computer tutor, we implemented two different feedback policies corresponding to two different interaction styles. In the fully interactive condition (full), the computer tutor (ITS) used a subset of the most frequent strategies we observed in the human-human corpus. When the student answered a question correctly, the ITS confirmed the correct answer, occasionally restating it with improved terminology [12]. When the student was incorrect, the ITS provided hints at two levels of specificity, or suggested slides to re-read. The system responded to negative metacognitive statements by giving hints, or giving away the complete answer if the student appeared to be having significant difficulties. It ignored positive metacognitive and social statements. Finally, the system said "I am sorry, I didn't understand" in response to nonsense, and remediated by providing hints depending on the tutoring context [26].

An example of the interaction with the system in the full condition is shown in Figure 2(a). In their first turn, the student gives an answer which is partially correct, but incomplete: the student did not mention that a battery must be contained in the same closed path. The answer is analyzed with a natural language parser and semantic interpreter, and the diagnoser determines that the student correctly mentioned a bulb in a closed path, but is missing the relationship with the battery. The tutor then decides that, given that this is

(a)
Tutor: What are the conditions for a bulb to light?
Student: It has to be in a closed path
Tutor: Right. The bulb must be in a closed path. Keep going.
Student: The battery must also be in a closed path.
Tutor: Right. A bulb must be contained in a closed path with the battery. When you are ready, go on to the next slide.

(b)
Tutor: What are the conditions for a bulb to light?
Student: It has to be in a closed path
Tutor: OK. One good way to phrase the answer is: A bulb must be contained in a closed path with the battery. When you are ready, go on to the next slide.

Fig. 2. Example interaction with the system a) in FULL condition; b) in BASE condition

the first student error for this question, the appropriate tutoring strategy is to re-state the correct part of the answer, and give a content-less prompt for the missing information.[6] The second student response completes the answer, but, since the answer was collected over multiple turns, the tutor restates it again as a complete sentence.

In contrast, in the baseline minimally-interactive condition (BASE), the ITS asks exactly the same questions, but does not provide specific feedback. Instead, it provides a neutral acknowledgment of the student's contribution, and gives away the correct answer. An example interaction is shown in Figure 2(b). As can be seen in this answer, the system provides the student with the correct answer, but makes no attempt to either acknowledge the correct part or point out specific problems. This condition is effectively equivalent to a non-interactive e-learning environment where students were asked to "write in" their answers and compare them with the system's answers. Students were not told that their answers weren't checked, in order to encourage them to provide meaningful answers.

Using the two conditions allows us to compare which factors correlated with learning gain and user satisfaction depending on the interaction style within a computer-based learning environment.[7]

4.2 Procedure

The procedure for the human-computer study was essentially the same as the human-human study with a few exceptions. The pre-test consisted of 22 multiple choice questions and the post-test consisted of 21 multiple choice questions. The

[6] If a student was performing poorly, more specific hints would be used, for example "Here's a hint: your answer should mention a battery."

[7] Our earlier study [25] examined data from FULL only, and served as a motivation for this extended analysis comparing all 3 conditions with additional variables of interest.

same set of questions was used in the human-human and human-computer studies.[8] The participants were also given a usability and satisfaction questionnaire developed to measure their satisfaction with different aspects of the system.

4.3 Corpus

Seventy six undergraduate students without prior knowledge of the domain from the same university as the human-human study were paid for participating in the human-computer study. The average age of the participants was 21.05 years ($SD = 3.30$) and there were almost twice as many females as males. There were 37 participants in the FULL and 39 participants in the BASE condition. The interaction logs were converted into the same XML format as the human-human corpus. The corpus includes an estimated 57,600 total dialogue turns taken by the student and tutor, and an estimated 680,000 words.

5 Data Analysis

The human-human and human-computer tutoring studies were conducted at different times, however, as we discussed in Section 4, they were using comparable learning environments. Therefore, for purposes of this study we conducted a three-way comparison between three different conditions found in our data: human-human interaction, human-computer dialogue with detailed feedback in FULL, and minimally interactive human-computer dialogue in BASE. All data was annotated using the same coding scheme, to compare contentful and non-contentful student statements.

5.1 Coding

The coding scheme we used is presented in Table 1 (reproduced from [25]). All student utterances were classified as primarily content, metacognitive, social or nonsense. Note that nonsense is a special category for dialogue with computers, defined as statements that are made up of random letters or numbers that are not content related (e.g., "ufghp"). It never occurred in dialogue with humans. For purposes of data analysis, we treated nonsense as instance of negative social, because it often appeared to be an expression of student frustration with the system, similar in function to expletives. For the human-human data, two independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, and social dialogue statements with a very high reliability (Cohen's kappa, $\kappa = 1.00$). In addition,

[8] The human-computer pre-test had fewer questions because in the human-human study, after taking the post-test the participants returned for a follow-up session with additional material. The pre-test covered the whole range of topics tutored by human tutors. In the human-computer study, only the first tutoring session was replicated by the ITS, and the pre-test questions were restricted to the material covered by the computer system and the first post-test.

Code	Definition	Example
Content	Statements including domain concepts that pertain to the lesson	"There is a battery and bulb in circuit 1." "1.5 volts."
Management	Dialogue that does not contain information relevant to the lesson material, but deals with the flow of the lesson	"I give up." "O.k." - Acknowledging the tutor's instructions to continue
Metacognition	Statements containing the student's feelings about his or her understanding, but does not include domain concepts	Metacognitive statements can be positive or negative.
Positive	Statements that express understanding	"I get it." "Oh, o.k."
Negative	Statements that express confusion	"I don't understand."
Social Dialogue	Dialogue that is not related to the content of the lessons or state of student's understanding, and expresses some form of social or emotional connection	Social statements can be positive or negative.
Positive	Statements that include humor, rapport, chit-chat, or saving face	"Ha-ha" "Hi, how are you doing?"
Negative	Statements that include frustration, refusal to cooperate with the system, or offending the system	"Because I said so." "No." "You're stupid." Expletives
Nonsense	Random sequences of letters or numbers	"oidhf" "dsfafadgdfh"

Table 1. Coding Summary

raters were able to differentiate between positive and negative metacognitive statements made by the student with high inter-rater reliability ($\kappa = 0.99$).

For the human-computer data, four independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, social dialogue, and nonsensical statements with high reliability ($\kappa = 0.88$), and to reliably differentiate between positive and negative metacognitive statements made by the student ($\kappa = 0.96$).

5.2 Variables of Interest

We analyzed and compared several properties of dialogue based on this coding. To investigate social and metacognitive aspects of the dialogue, we computed correlations between the learning gain and the number of metacognitive and social (including nonsense) statements submitted by the student during the course of each tutoring session. Management was left out of the analyses because it was not very prevalent in the computer tutoring data and in the few cases it did occur, it was ignored by the tutor. Also, it was not a relevant predictor of learning gain with the human tutor.

As mentioned in Section 2, we are also interested in the percentage of contentful talk, since it is known to be correlated with learning gain. There are multiple ways to define contentful talk [15]. For purposes of this study, we defined "content" as the number of student words present in a glossary compiled for the domain by instructional designers, normalized by the total number of words in the dialogue. We used the same glossary as [15].

6 Results

An initial study comparing 41 students in FULL to human tutors was presented in [25]. It demonstrated that there were significant differences in the distribution of metacognitive and social statements between FULL and human-human tutoring, as well as differences in which variables were correlated in learning gain. This study is extending results to compare different interaction styles, and the impact of contentful talk. However, in a small number of cases (4 participants) the data logs were incomplete due to technical failures, and did not contain the information necessary to compute measures of contentful talk. For consistency, we removed those participants from the set and replicated all our analyses, comparing measures of contentful talk, metacognitive, and social statements for the same set of participants.

6.1 Learning Gain

Pre and post-test scores were calculated in terms of percentage correct. A learning gain score normalized for pre-test performance was then calculated for each participant using the formula $gain = (posttest - pretest)/(1 - pretest)$.

6.2 Content-Learning Correlations

The summary of language statistics in the different conditions is shown in Table 2. As can be seen from the table, the numbers of words and turns in each session differed between conditions. This difference is due, in part, to splitting multi-part questions into individual questions (as discussed in Section 4). However, comparing average turn length and the percentage of contentful talk allows us to observe the differences in the student's dialogue behavior which are less affected by the differences in overall number of questions.

In human-human dialogue, students on average produced 5.68 words per turn ($SD = 2.22$). Overall, the percentage of content words produced by students (out of all words in the corpus) was significantly correlated with learning gain ($r = 0.40$, $p = 0.02$).

In human-computer dialogue, there were significant differences in terms of dialogue length and content between BASE and FULL (Student's t-test, $t(38) = 15.99, p < 0.0001$) However, students produced on average turns of similar length in the two conditions (Student's t-test, $t(71) = 1.61$, $p = 0.11$) and similar percentage of contentful words ($t(73) = 1.3$, $p = 0.20$). The proportion of contentful

Condition	Turns per session	Words per session	Words per turn	Content words per session	% content
Human-Human	144(51.8)	816(357.3)	5.68(2.22)	381(173.8)	42.1(5.6)
BASE	156(6.3)	726(224.2)	6.13(0.68)	455 (108.3)	51.3 (1.2)
FULL	232(33.8)	1411(219.4)	5.66(0.57)	741 (102.5)	52.5 (1.1)

Table 2. Student language statistics for human-human and human-computer dialogue (Standard deviations in parentheses)

Condition	Metacognitive		Social	
	Positive	Negative	Positive	Negative
Human	12.5 (8.16)	1.77 (1.94)	5.83 (6.99)	0
BASE	0	1.44 (2.93)	0.10 (0.68)	0.23 (0.84)
FULL	0.21 (0.48)	5.65 (4.24)	0.14 (0.54)	4.21 (8.03)

Table 3. Mean number of metacognitive and social statements per session in different conditions (standard deviation in parentheses).

words out of all words in the dialogue was significantly higher than in human-human dialogue ($t(38) = 9.05$, $p < 0.0001$). This reflects a general difference between the human-human and human-computer interaction styles. The interaction style of the computer tutor discouraged social comments by not reacting to them.

Even though students produced the same percentage of contentful talk in the two human-computer conditions, the proportion of contentful talk was only correlated with learning gain in FULL ($r = 0.42$, $p = 0.009$). In BASE, the proportion of contentful talk was not correlated with learning gain ($r = 0.19$, $p = 0.25$). We discuss possible reasons for this in Section 7.

6.3 Metacognitive Statements

The number of occurrences of metacognitive and social statements is summarized in Table 3. Students made metacognitive statements in all conditions, regardless of whether the tutor was a human or a computer; however the relative frequencies of positive and negative metacognitive statements depended on the type of tutor.

Students in FULL made significantly more metacognitive statements than students in BASE ($t(53) = 2.84$, $p < 0.01$). For either condition, this was significantly smaller than the number of metacognitive statements in human-human dialogue (FULL: $t(31) = -5.98$, $p < 0.001$; BASE: $t(29) = -6.58$, $p < 0.001$).

Students talking to a human tutor made significantly more positive metacognitive statements than negative metacognitive statements (paired t-test, $t(29) = 8.37$, $p < 0.001$). In contrast, students talking to both computer tutors, made significantly more negative metacognitive statements than positive metacognitive statements (FULL: $t(36) = -4.42$, $p < 0.001$; BASE: $t(38) = -3.05$, $p < 0.01$).

The condition also affected how the metacognitive statements were related to learning gain. For students interacting with a human tutor, the amount of positive metacognitive dialogue was significantly negatively correlated with learning gain ($r = -0.543$, $p = .002$), while the number of negative metacognitive statements was not correlated with learning gain ($r = -0.210$, $p = 0.266$). For students interacting with BASE, there were no positive metacognitive statements, and negative metacognitive statements were not correlated with learning gain ($r = -0.19$, $p = 0.25$). Finally, for students interacting with FULL, the number of both types of metacognitive statements were significantly negatively correlated with learning gain (positive statements: $r = -0.419$, $p = 0.006$; negative statements: $r = -0.537$, $p < .001$).

6.4 Social Statements and Nonsense

While students made social statements with both types of tutors, students interacting with a human tutor made exclusively positive social statements and students interacting with the computer tutor made exclusively negative social statements, either negative comments or submitting nonsense.

Again, students in FULL made significantly more social statements than students in BASE ($t(37) = 2.56$, $p = 0.01$). The overall number of social statements made with the computer tutor was significantly lower than in human-human dialogue (FULL: $t(30) = -2.80$, $p = 0.009$; BASE: $t(29) = -3.23$, $p = 0.003$).

In the human-human condition, the amount of social dialogue was not significantly correlated with learning gain. However, in FULL the average learning gain score of the students who generated any negative social dialogue, 52% ($SD = 26$), was statistically significantly lower than the average learning gain score, 67% ($SD = 12$), of students who did not ($t(38) = -2.43$, $p = 0.02$). Not surprisingly, the amount of negative social dialogue generated by the students in this condition was also significantly negatively correlated with the students' report of satisfaction with the computer tutor, $r = -0.55, p < 0.001$. There were too few instances of negative social in BASE to produce any meaningful statistical results.

6.5 The impact of interpretation failures

While the FULL system attempts to model interaction with the human by accepting extended answers to "Why" questions and giving extended feedback, there are still significant differences in the system's capabilities compared to human-human interaction, due to the limitations of the natural language technology. Overall, the system failed to interpret 13.4% of student utterances ($SD = 5.00$). The frequency of interpretation problems was significantly negatively correlated with learning gain ($r = -0.47, p < 0.005$) and with user satisfaction ($r = -0.36$, $p < 0.05$). The interpretation problems encountered by students therefore present a major challenge for system implementation, and for comparison with human-human interaction. We discuss the implications in the next section.

7 Discussion and Future Work

As previously mentioned, it is common for ITSs to be modeled after human tutors, but it is uncertain if this is an effective technique because we are unsure if these interactions are similar and can be interpreted in the same way. The goal of developing the Beetle II system was to provide a tutor that can accept a wider range of user language than existing tutors, and therefore deliver interaction style more similar to human tutoring.

While we have not completely achieved this goal (as evidenced by the impact of interpretation problems), the data we collected provide a new and significant comparison between human-human and human-computer interaction. Previously available studies compared human-human dialogue with a system that was only capable of asking short-answer questions during dialogue [27, 23]. In these studies, the student's response when interacting with a human was on average 3 times longer than when interacting with a computer. Our system successfully interpreted a range of longer sentences, and the average turn length was more similar in human-human and human-computer tutoring.

Even as average turn length increased, factors correlated with learning gain also depended on the interaction style. In both human-human and full condition, the percentage of contentful talk was correlated with learning gain. This relationship did not hold for the base condition where students were not receiving targeted feedback. One possible explanation is that students often struggle with using appropriate terminology in this domain. At least one participant commented during the post-lesson interview that they found it difficult to figure out whether the difference between their answer and the answer given by the system was important or trivial. This indicates that it may not be enough to prompt students to explain their answers in an e-learning environment. The system actually needs to give them feedback that targets specific problems in their answer in order to "convert" the increases in contentful talk into learning gains.

The differences in interaction style had an impact on social and metacognitive dialogue as well. With a human tutor, the non-content statements were mostly positive acknowledgments and social statements used to build rapport; with computer tutors, students mostly used negative statements expressing confusion or showing frustration with the system's inability to interpret the user's input correctly or generate appropriate feedback.

We previously concluded that the differences in behavior between human-human and human-computer tutoring indicate that the negative social and metacognitive statements are more reliable indicators of student frustration in human-computer dialogue than in human-human dialogue, and therefore need to be addressed especially [25]. Adding base to this analysis introduces a new dimension, namely, interaction style. It is clear that students reacted with more social and metacognitive statements when the tutor (either human or computer) was listening and responding adaptively, while the neutral responses of base did not elicit either social or metacognitive statements.

An open question with respect to this analysis is to which extent the negative social and metacognitive statements are influenced by the limitations in the system's interpretation capabilities. Students used fewer negative social statements in BASE, where the system never indicated that they were not understood. To some extent frustration can be mitigated by improving language understanding components, thus reducing the number of misunderstandings. However, this may not be sufficient by itself. The negative correlation between learning gain and interpretation errors was observed in both BASE and FULL, despite students being more satisfied with BASE. These differences are further investigated in [28], where we conclude that special strategies are necessary for dealing with incorrect or vague use of student terminology.

Moreover, the role of motivational factors needs to be further examined. Previous research found that students' attitudes towards learning and expectations of computer systems affected their frustration and learning gain. Students who, before any interaction with the system, didn't believe that a computer system can help them learn were more frustrated with a computer tutor, even though they learned as much as the students who believed that a computer tutor would be helpful [29]. In a case-based learning environment, novice students didn't see the need to write down their explanations, therefore skipping this step when given an option. However, students who were forced to produce explanations learned more, even though they were less happy with the system [30]. Thus, reducing student frustration may not always have a positive effect on learning, and designers should focus on determining which negative metactognitive and social expressions indicate that learning is negatively affected and require action from the system. There is now some work in detecting and responding to student uncertainty in human-computer dialogue [31]. Future work should also investigate how students' prior attitudes to learning with computers affect their frustration with interpretation problems, and how to devise better error-recovery strategies to deal with the unavoidable limitations of natural language technology.

8 Conclusion

In this study we compared human-human tutorial dialogue with two different computer tutoring styles, focusing on features correlated with learning gain. We found significant differences between human-human and human-computer tutoring, but also in how students interact with different computer tutors teaching the same material. This could be partially explained by the limitations in computer system capabilities and frustration arising from problems in system interpretation. However, student attitudes to computers and learning may play an important role as well. Moreover, different factors are associated with learning gain depending on the system interaction style. Our results indicate that giving adaptive feedback changes which factors are correlated with learning gain compared to a system that simply gives away the answers. Further research should particularly focus on dealing with negative social and metacognitive statements, and address student beliefs which may be causing frustration, in addition to improving the system's ability to correctly interpret student's answers.

References

1. Aleven, V., Popescu, O., Koedinger, K.R.: Towards tutorial dialog to support self-explanation: Adding natural language understanding to a cognitive tutor. In: Proceedings of the 10^{th} International Conference on Artificial Intelligence in Education (AIED '01)". (2001)
2. Goguadze, G., González Palomo, A., Melis, E.: Interactivity of exercises in activemath. Towards Sustainable and Scalable Educational Innovations Informed by the Learning Sciences Sharing. Good Practices of Research Experimentation and Innovation. **133** (2005) 109–115
3. Bloom, B.S.: The two sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher **13** (1984) 3–16
4. Rosé, C., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., Weinstein, A.: Interactive conceptual tutoring in atlas-andes. In: Proceedings of AI in Education 2001 Conference. (2001)
5. VanLehn, K., Jordan, P., Litman., D.: Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In: Proceedings of SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA (October 2007)
6. Aleven, V., Popescu, O., Koedinger, K.: Pilot-testing a tutorial dialogue system that supports self-explanation. Lecture Notes in Computer Science **2363** (2002) 344–354
7. Pon-Barry, H., Clark, B., Schultz, K., Bratt, E.O., Peters, S.: Advantages of spoken language interaction in dialogue-based intelligent tutoring systems. In: Proceedings of ITS-2004. (2004) 390–400
8. Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, P., Kreuz, R.: Autotutor: A simulation of a human tutor. Cognitive Systems Research **1** (1999) 35–51
9. Callaway, C., Dzikovska, M., Matheson, C., Moore, J., Zinn, C.: Using dialogue to learn math in the LeActiveMath project. In: Proceedings of the ECAI Workshop on Language-Enhanced Educational Technology. (August 2006) 1–8
10. Khuwaja, R.A., Evens, M.W., Michael, J.A., Rovick, A.A.: Architecture of CIRCSIM-tutor (v.3): A smart cardiovascular physiology tutor. In: Proceedings of the 7th Annual IEEE Computer-Based Medical Systems Symposium. (1994)
11. Nielsen, R.D., Ward, W., Martin, J.H.: Learning to assess low-level conceptual understanding. In: Proceedings 21st International FLAIRS Conference, Coconut Grove, Florida (May 2008)
12. Dzikovska, M.O., Campbell, G.E., Callaway, C.B., Steinhauser, N.B., Farrow, E., Moore, J.D., Butler, L.A., Matheson, C.: Diagnosing natural language answers to support adaptive tutoring. In: Proceedings 21st International FLAIRS Conference, Coconut Grove, Florida (May 2008)
13. Jordan, P.W.: Using student explanations as models for adapting tutorial dialogue. In Barr, V., Markov, Z., eds.: FLAIRS Conference, AAAI Press (2004)
14. Purandare, A., Litman, D.: Content-learning correlations in spoken tutoring dialogs at word, turn and discourse levels. In: Proceedings 21st International FLAIRS Conference, Coconut Grove, Florida (May 2008)
15. Litman, D., Moore, J., Dzikovska, M., Farrow, E.: Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. In: Proceedings of 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK (July 2009)
16. Campbell, G.C., Steinhauser, N.B., Dzikovska, M.O., Moore, J.D., Callaway, C.B., Farrow, E.: Metacognitive awareness versus linguistic politeness: Expressions of

confusions in tutorial dialogues. In: Poster presented at the 31st Annual Conference of the Cognitive Science Society, Amsterdam, Netherlands (July 2009)

17. Reeves, B., Nass, C.: The media equation: how people treat computers, television, and new media like real people and places. Cambridge University Press, New York, NY, USA (1996)
18. Zanbaka, C., Ulinski, A., Goolkasian, P., Hodges, L.F.: Effects of virtual human presence on task performance. In: Proceedings of International Conference on Artificial Reality and Telexistence (ICAT). (2004) 174–181
19. Pertaub, D.P., Slater, M., Barker, C.: An experiment on public speaking anxiety in response to three different types of virtual audience. Presence: Teleoper. Virtual Environ. **11**(1) (2002) 68–78
20. Shechtman, N., Horowitz, L.M.: Media inequality in conversation: how people behave differently when interacting with computers and people. In: CHI'03: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (2003) 281–288
21. Rosé, C., Torrey, C.: Interactivity versus expectation: Eliciting learning oriented behavior with tutorial dialogue systems. In: Proceedings of Interact'05. (2005)
22. Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C.: Eliciting self-explanations improves understanding. Cognitive Science **18**(3) (1994) 439–477
23. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rosé, C.P.: When are tutorial dialogues more effective than reading? Cognitive Science **31**(1) (2007) 3–62
24. Bransford, J.D., Brown, A.L., Cocking, R.R., eds.: How People Learn: Brain, Mind, Experience, and School Committee on Developments in the Science of Learning. Commission on Behavioral and Social Sciences and Education of the National Research Council, National Academy Press (2000)
25. Steinhauser, N.B., Campbell, G.E., Harrison, K.M., Taylor, L.S., Dzikovska, M.O., Moore, J.D.: Comparing human-human and human-computer tutorial dialogue. In: Proceedings of the 32nd Annual Conference of the Cognitive Science Society poster session. (2010)
26. Dzikovska, M.O., Callaway, C.B., Farrow, E., Moore, J.D., Steinhauser, N.B., Campbell, G.C.: Dealing with interpretation errors in tutorial dialogue. In: Proceedings of SIGDIAL-09, London, UK (Sep 2009)
27. Litman, D., Rosé, C.P., Forbes-Riley, K., VanLehn, K., Bhembe, D., Silliman, S.: Spoken versus typed human and computer dialogue tutoring. International Journal of Artificial Intelligence in Education **16** (2006) 145–170
28. Dzikovska, M.O., Moore, J.D., Steinhauser, N., Campbell, G.: The impact of interpretation problems on tutorial dialogue. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics(ACL-2010). (2010)
29. Jackson, G.T., Graesser, A.C., McNamara, D.S.: What students expect may have more impact than what they know or feel. In: Proceedings 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK (2009)
30. Papadopoulos, P.M., Demetriadis, S.N., Stamelos, I.: The impact of prompting in technology-enhanced learning as moderated by students' motivation and metacognitive skills. In: Proceedings of 4th European Conference on Technology Enhanced Learning (EC-TEL 2009). (2009)
31. Forbes-Riley, K., Litman, D.: Adapting to student uncertainty improves tutoring dialogues. In: Proceedings 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK (2009)